The Binomial Distribution

August 14, 2019

Stem and Leaf Plots

stem	leaf
0	1, 1, 2, 2, 3, 4, 4, 4, 4, 5, 8
1	0, 0, 0, 1, 1, 3, 7, 9
2	5, 5, 7, 7, 8, 8, 9, 9
3	0, 1, 1, 1, 2, 2, 2, 4, 5
4	0, 4, 8, 9
5	2, 6, 7, 7, 8
6	3, 6

Key: 6|3 = 63 years old

Example: Insurance Deducibles

- Suppose a health insurance company found that 70% of the people they insure stay below their deductible in any given year.
- Each of these people can be thought of as a single trial in a study.
- We label a person a "success" if their healthcare costs do not exceed the deductible.
  - $P(\texttt{success}) = p = 0.7$
  - $P(\texttt{failure}) = 1 - p = 0.3$

The Bernoulli Distribution

- When an individual trial only has two possible outcomes it is called a Bernoulli random variable.
  - These outcomes are often labeled as success or failure.
- *These labels can be completely arbitrary!*
  - We called "not hitting the deductible" a "success", but we could just as well have labeled that the "failure".
  - The framework we use to talk about the Bernoulli distribution does not depend on the label we use.

The Bernoulli Distribution

Bernoulli random variables are often denoted as `1` for a success and `0` for a failure.

This makes data entry easy and is mathematically convenient.

Suppose we observe ten trials:

$$1, \quad 1, \quad 1, \quad 0, \quad 1, \quad 0, \quad 0, \quad 1, \quad 1, \quad 0$$

The Sample Proportion

The **sample proportion**, $\hat{p}$, will be the sample mean for these observations:

$$\hat{p} = \frac{\#\text{ of successes}}{\#\text{ of trials}}$$

$$= \frac{1+1+1+0+1+0+0+1+1+0}{10}$$

$$= 0.6$$

The Bernoulli Random Variable

- It is useful to think about a Bernoulli random variable as a random process with only two outcomes: a success or failure (or yes/no).
- Then we code a success as 1 and a failure as 0.
- These are just numbers, so we can define the mean and variance.

The Bernoulli Random Variable

If X is a random variable that takes the value 1 with probability of success p and 0 with probability $1-p$, then X is a **Bernoulli random variable** with mean

$$\mu = p$$

and variance

$$\sigma^2 = p(1-p).$$

The Bernoulli Distribution

- Remember that we can estimate p using $\hat{p} = \bar{x}$.
- We can use this to estimate the mean the variance.
- For our insurance deductible example, we found $\hat{p} = 0.6$
- So we can estimate

$$\hat{\mu} = \hat{p} = 0.6$$

and

$$\hat{\sigma^2} = \hat{p}(1 - \hat{p}) = 0.6 * 0.4 = 0.24$$

Example

Derive the mean and variance of a Bernoulli random variable.

Example

Because there are only 2 possible outcomes, the Bernoulli distribution describes a *discrete random variable.*

Therefore, We can start with its probability distribution table:

x	0	1
$P(x)$	p	$(1-p)$

Example

Then for the expected value,

x	1	0	Total
$P(x)$	p	$(1-p)$	
$xP(x)$	p	0	p

So the expected value is (as expected) p!

Example

And for the variance,

x	1	0	Total
$P(x)$	p	$(1-p)$	
$xP(x)$	p	0	p
$x-E(x)$	$1-p$	$-p$	
$[x-E(x)]^2$	$(1-p)^2$	p^2	
$P(x)[x-E(x)]^2$	$p(1-p)^2$	$(1-p)p^2$	$p(1-p)^2+(1-p)p^2$

Example

Then

$$\begin{aligned} Var(X) &= p(1-p)^2 + (1-p)p^2 \\ &= p - 2p^2 + p^3 + p^2 - p^3 \\ &= p - 2p^2 + p^2 \\ &= p - p^2 \\ &= p(1-p) \end{aligned}$$

Which is the Var(X) we wanted!

The Binomial Distribution

The **binomial distribution** is used to describe the number of successes in a fixed number of trials.

- This is an extension of the Bernoulli distribution.
- We check for a success or failure repeatedly over multiple trials.
- Each *individual* trial can be described with a Bernoulli distribution.

Example: Insurance

- Let's return to the insurance agency where 70% of individuals do not exceed their deductible.
- Suppose the insurance agency is considering a random sample of four individuals they insure.
- What is the probability that exactly one of them will exceed the deductible and the other three will not?

Example

Let's call the four people Ariana (A), Brittany (B), Carlton (C), and Damian (D). Consider a scenario where one person exceeds the deductible:

$$\begin{aligned}
&P(A = \texttt{exceed}, B = \texttt{not}, C = \texttt{not}, D = \texttt{not}) \\
&\quad = P(A = \texttt{exceed}) \times P(B = \texttt{not}) \times P(C = \texttt{not}) \times P(D = \texttt{not}) \\
&\quad = (0.3) \times (0.7) \times (0.7) \times (0.7) \\
&\quad = (0.3)^1 \times (0.7)^3 \\
&\quad = 0.103
\end{aligned}$$

Example

- But there are three other scenarios!
  1. Brittany could have been the one to exceed the deductible.
  2. ... or Carlton could have.
  3. ... or Damian.
- In each of these cases, the probability is $(0.7)^3(0.3)^1$.

Example

- These four scenarios consist of all the possible ways that exactly one of these four people could have exceeded the deductible.
- So the total probability is

$$4 \times (0.7)^3 \times (0.3)^1 = 0.412.$$

This is an example of a scenario where we would use a binomial distribution.

The Binomial Distribution

We would like to determine the probabilities associated with the binomial distribution using n, k, and p.

We would like a nice formula for this.

Example: Building to Binomial

Let's return to our insurance example.

- There were four people who could have been the single failure.
- Each scenario has the same probability.
- So the final probability was

$$[\# \text{ of scenarios}] \times P(\text{single scenario})$$

Example: Building to Binomial

- The first component of this equation is the number of ways to arrange $k = 3$ successes among $n = 4$ trials.
- The second is the probability of any one of the scenarios.
  - These four scenarios are equally probable.

Building to Binomial

- Consider P(single scenario) with k successes and $n - k$ failures in n trials.
- We know how to handle this!
- We will use the multiplication rule for independent events.

Probability for a Single Scenario

Applying the multiplication rule for independent events,

$$\begin{aligned} P(\text{single scenario}) &= P(k \text{ successes}) \times P(n-k \text{ failures}) \\ &= p \times \cdots \times p \times (1-p) \times \cdots \times (1-p) \\ &= p^k \times (1-p)^{n-k} \end{aligned}$$

This is our general formula for $P(\text{single scenario})$.

Number of Ways to Arrange Successes

The number of ways to arrange k successes and $n-k$ failures is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

The expression $\binom{n}{k}$ is read "n choose k". This is the number of ways to *choose* k successes in n trials.

What about the exclamation point?

Factorial Notation

The exclamation point in $n!$ denotes a **factorial**.

$$
\begin{aligned}
0! &= 1 \\
1! &= 1 \\
2! &= 2 \times 1 \\
3! &= 3 \times 2 \times 1 \\
4! &= 4 \times 3 \times 2 \times 1 \\
&\vdots \\
n! &= n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1
\end{aligned}
$$

Example

We can use this to double check our insurance deductible problem.

Recall that we decided that there were four possible ways to get 3 successes (not exceeding) among 4 people (trials).

$$\begin{aligned}\binom{4}{3} &= \frac{4!}{3!(4-3)!} \\ &= \frac{4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1) \times (1)} \\ &= 4\end{aligned}$$

which is just what we decided before!

The Binomial Distribution

Suppose $X \sim \text{Bin}(n, p)$. The probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

The Binomial Distribution

The expected value (mean) is

$$E(X) = \mu = np$$

and the variance is

$$Var(X) = \sigma^2 = np(1-p)$$

If $p \approx (1-p)$, then the binomial distribution is symmetric.

The Binomial Distribution

We say that X follows a **binomial distribution** with number of trials n and probability of success p if

1. The number of trials is fixed $= n$.
2. The trials are independent.
3. There are two possible outcomes, success/failure.
4. The probability of success is known and fixed $= p$.

We denote this $X \sim \text{Bin}(n, p)$

Example: Cars at UCR

In a survey conducted at UCR, it is reported that 38% of students owned a car. A random sample of 20 STAT 100A students is selected. Let X be the number of students in the sample who own a car. What is the distribution of X?

Example: Cars at UCR

In a survey conducted at UCR, it is reported that 38% of students owned a car. A random sample of 20 STAT 100A students is selected. Let X be the number of students in the sample who own a car. What is the distribution of X?

1. $n = 20$ students, so the number of trials is fixed.
2. We have a random sample, so the trials are independent.
3. Success = `car`
   Failure = `no car`
4. $p = P(\texttt{car}) = 0.38$

So $X \sim \text{Bin}(n = 20, p = 0.38)$

Example: Cars at UCR

What is the probability that none of the 20 students own a car?

Example: Cars at UCR

What are the mean and variance of X, the number of students in the sample who own a car?

Computing Binomial Probabilities

1. Check that the (binomial)model is appropriate.
2. Identify n, p, and k.
3. Determine the probability.
4. Interpret the results.

When doing calculations by hand, cancel out as many terms as possible in the binomial coefficient!

Example: Cars at UCR

What is the probability that no more than 2 students own a car?

Example: Cars at UCR

What is the probability that fewer than two students own a car?

Example: Cars at UCR

What is the probability that more than 2 students own a car?

Normal Approximation to the Binomial Distribution

- Sometimes when n is large, the binomial formula can be difficult to use.
- In these cases, we may be able to use the normal distribution to estimate binomial probabilities.

Example

- Approximately 15% of the US population smokes cigarettes.
- A local government commissioned a survey of 400 randomly selected individuals.
- The survey found that only 42 of the 400 participants smoke cigarettes.
- If the true proportion of smokers in the community was really 15%, what is the probability of observing 42 or fewer smokers in a sample of 400 people?

Example

First, we check that this is a binomial setting:

1. $n = 400$ community members
2. This is a random sample, so the trials are independent.
3. We define Success = `smoker` and Failure = `nonsmoker`.
4. $p = P(\texttt{smoker}) = 0.15$

So this is a binomial distribution.

We are interested in $k = 42$ *or fewer*.

Example

Let X be the number of smokers in a community. We want to know

$$P(X \leq 42)$$

which is the same as

$$\begin{aligned} &P(X = 42 \text{ or } X = 41 \text{ or } X = 40 \text{ or } \ \dots \ \text{ or } X = 1 \text{ or } X = 0) \\ &= P(X = 42) + P(X = 41) + \cdots + P(X = 1) + P(X = 0) \end{aligned}$$

We *could* calculate each of the 43 probabilities individually by using our binomial formula and adding them together...

Example

If we were to do this, we would find

$$P(X = 42) + P(X = 41) + \cdots + P(X = 1) + P(X = 0) = 0.0054$$

That is, if the true proportion of smokers in the community is $p = 0.15$, then the probability of observing 42 or fewer smokers in a sample of $n = 400$ is 0.0054.

Normal Approximation to the Binomial Distribution

...but why would we do this if we don't have to?

- Calculating probabilities for a range of values is much easier using the normal model.
- We'd like to use the normal model in place of the binomial distribution.

Normal Approximation to the Binomial Distribution

Surprisingly, this works quite well as long as

$$np > 10$$

and

$$n(1 - p) > 10$$

Note that *both of these conditions must hold*!

Normal Approximation to the Binomial Distribution

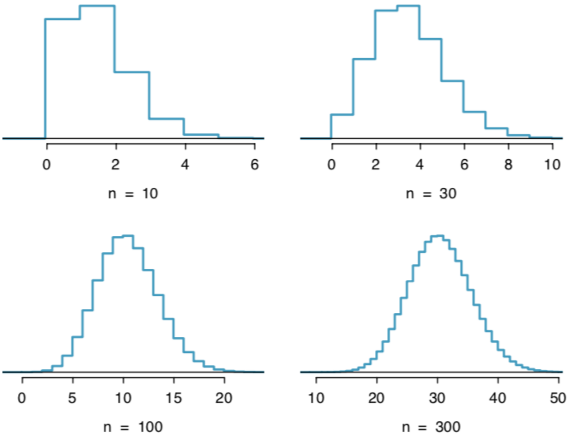
If these conditions are met, then $X \sim \text{Bin}(n, p)$ is well-approximated by a normal model with

$$E(X) = \mu = np$$

and

$$Var(X) = \sigma^2 = np(1 - p).$$

Normal Approximation to the Binomial Distribution



Each histogram shows a binomial distribution with $p = 0.1$.

Example

Can we use the normal approximation to estimate the probability of observing 42 or fewer smokers in a sample of 400, if the true proportion of smokers is $p = 0.15$?

Example

Can we use the normal approximation to estimate the probability of observing 42 or fewer smokers in a sample of 400, if the true proportion of smokers is $p = 0.15$?

From our previous example, we verified that the binomial model is reasonable. Now,

$$np = 400 \times 0.15 = 60$$

and

$$n(1-p) = 400 \times 0.85 = 340$$

so both are at least 10 and we may use the normal approximation.

Example

For the normal approximation,

$$\mu = np = 400 \times 0.15 = 60$$

and

$$\sigma = \sqrt{np(1-p)} = \sqrt{400 \times 0.15 \times 0.85} = 7.14$$

Example

We want to find the probability of observing 42 or fewer smokers using or $N(\mu = 60, \sigma = 7.14)$ model.

We start by finding our Z-score:

$$z = \frac{x - \mu}{\sigma} = \frac{42 - 60}{7.14} = -2.52$$

Example

- Then, using `R`, the left-tail area is 0.0059.
- When we calculated this using the binomial distribution, the true probability was 0.0054.
- So this is a pretty good approximation!

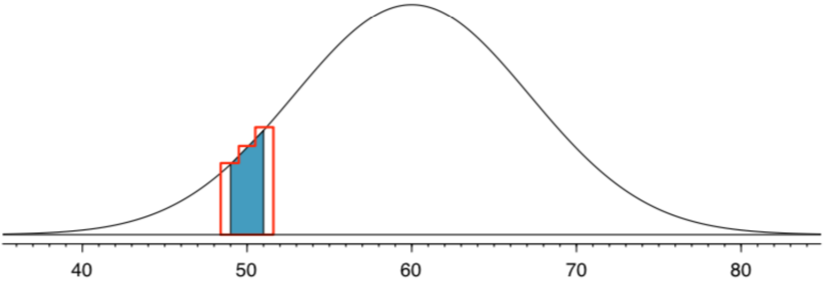
Breakdown of the Normal Approximation

- The normal approximation to the binomial distribution tends to perform poorly when estimating the probability of a small range of counts.
- This is true even when $np > 10$ and $n(1-p) > 10$

Breakdown of the Normal Approximation

- Suppose we wanted to compute the probability of observing 49, 50, or 51 smokers in 400 when $p = 0.15$.
- We know that $np = 60 > 10$ and $n(1-p) = 340$, so we might want to apply the normal approximation and use the range 49 to 51.
- But this time the approximation and the binomial solution are noticeably different!
  - Binomial: 0.0649
  - Normal: 0.0421

Why Does This Breakdown Happen?



The binomial probability is shown outlined in red; the normal probability shaded in blue.

Can We Fix It? Improving the Normal Approximation for Intervals

We can usually improve this estimation by modifying our cutoff values.

- Cutoff values for the left side should be reduced by 0.5.
- Cutoff values for the right side should be increased by 0.5.

Example

- Suppose we wanted to compute the probability of observing 49, 50, or 51 smokers in 400 when $p = 0.15$.
- Let's try this again with our modification.
- For our normal distribution, we used a $N(60, 7.14)$ model.
- Our upper value is 51, adjusted to $51 + 0.5 = 51.5$.
- Our lower value is 49, adjusted to $49 - 0.5 = 48.5$.

Example

Then

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{51.5 - 60}{7.14} = -1.190476$$

and

$$z_2 = \frac{x_2 - \mu}{\sigma} = \frac{48.5 - 60}{7.14} = -1.610644$$

Example

Now, using `R`,

$$
\begin{aligned}
P(z_2 < Z < z_1) &= P(Z < z_1) - P(Z < z_2) \\
&= 0.1169297 - 0.05362867 \\
&= 0.0633
\end{aligned}
$$

Example

$P(49 \leq X \leq 51)$		
Binomial	Normal Approx (Adjusted)	Normal Approx (Unadjusted)
0.0649	0.0633	0.0421

Making those small adjustments makes a significant difference!